Discovery of Fur Binding Site Clusters in *Escherichia coli* by an Information Theory Model

Karen A. Lewis[*], [†]Zehua Chen[*], Ryan K. Shultzaberger[*], Ilya G. Lyakhov[‡], Ming Zheng[§,¶], Bernard Doan[‡,‖], Gisela Storz[‡], and Thomas D. Schneider[**]

version = 2.65 of fur.tex 2004 Jun 23

Fur regulates bacterial iron uptake systems. Twelve footprinted *Escherichia coli* Fur binding sites were used to create an information theory model of Fur binding. When the model was scanned across the twelve sequences, sequence walkers, which are visual depictions of predicted binding sites, frequently appeared in clusters that fit the published footprint data. This indicated that the model could accurately predict Fur binding. Within the clusters, individual walkers were separated from their neighbors by exactly 3 or 6 bases consistent with models in which Fur dimers either bind on opposite sides of the DNA helix or self-compete. When the *E. coli* genome was scanned, we found at least 40 other clusters. DNase I footprinting was used to examine purified Fur binding to the strongest site in the genome, in the *fhuF* promoter region. The footprints showed two distinct protected regions, and these were each covered by a cluster of walkers to within 6 base pairs. Gel mobility shift assays with 15 sites in the genome showed that the information theory model successfully predicts Fur binding sites and avoids sites that do not bind Fur even though they were predicted by a consensus sequence model.

Keywords: overlapping binding sites, *fhuF*, sequence walkers, Fur, information theory.

[*]National Cancer Institute at Frederick, Laboratory of Experimental and Computational Biology, P. O. Box B, Frederick, MD 21702-1201, USA. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598.

[†] **current address: University of Texas Southwestern Medical Center at Dallas, Department of Physiology, 5323 Harry Hines Blvd., Dallas, TX 75390-9040**

[‡]Basic Research Program, SAIC-Frederick, Inc National Cancer Institute at Frederick, Frederick, MD 21702-1201, USA

[§]National Institute of Child Health and Human Development, Cell Biology and Metabolism Branch, Building 18T, Room 101, Bethesda, MD 20892-5430, USA. (301) 402-0968, fax: (301) 402-0078.

[¶]current address: Dupont Central Research and Development, Experimental Station E328-B31, P.O. Box 80328, Wilmington, DE 19880-0328, USA. (302)-695-7136

[‖]Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, 12441 Parklawn Drive, Twinbrook II, Room 239, Rockville, MD 208522-1742

[**]Corresponding author. toms@ncifcrf.gov, http://www.lecb.ncifcrf.gov/~toms/

Introduction

The protein Fur is the 16.8 kDa product of the *fur* (*f*erric *u*ptake *r*egulation) gene in *Escherichia coli* (1), so named because it was first observed to repress the transcription of genes that code for components of ferric (Fe^{+3}) uptake systems found in the cell membrane. Since then, Fur also has been found to regulate other genes that are not directly related to iron transport, such as those encoding hemolysin, Shiga-like toxin, and manganese superoxide dismutase (2–5).

Fur binds to DNA and represses transcription in the presence of divalent metal ions. The ion is thought to be Fe^{+2} *in vivo* (6), however, DNase I footprinting experiments have shown that Fur also binds to DNA in the presence of Mn^{+2}, Co^{+2}, Cu^{+2}, Cd^{+2}, and Zn^{+2} (7). Recent studies have suggested that purified Fur contains at least one Zn^{+2} ion as a structural stabilizer (8). Fur has been observed to bind to DNA as a dimer and in higher order polymers (7), and electron microscopy has shown polymerization of Fur on DNA under high concentrations of protein and metal ions (2).

Numerous strategies have been employed to find new Fur binding sites. Various consensus sequences have been derived from both footprinted and non-footprinted Fur binding sites (3, 7, 9) and these have been compared to sequences in the promoter region of suspected iron-regulated genes. Putative Fur targets were then investigated further through genetic and biochemical experiments. Stojiljkovic *et al.* created a successful 'Fur titration assay' to locate new Fur binding sites using an *fhuF*:*lacZ* fusion and Fur consensus sequence-containing plasmid titrant on MacConkey plates (1). Several new iron-regulated genes in *E. coli* were discovered using this consensus sequence-based technique. In addition to the above, studies have also been carried out using *E. coli* Fur for DNase I footprinting with non-*E. coli* DNA (10,11). Transcriptional profiles of *E. coli* genes have also been used to determine those that are regulated by iron and Fur by evaluating mRNA levels in the absence of iron and Fur protein (12).

Another method for finding Fur-regulated genes is to use molecular information theory to locate new binding sites. Using this approach, classical information theory (13, 14) is applied to molecular biology (15). First, a set of binding sites is aligned by maximizing the information content (16), and then the average pattern at the sites is represented by a computer graphic called a sequence logo (17). Next, the conservation of bases in the aligned set is used to create a weight matrix model that assigns a weight in bits to each base at each position according to its frequency in the data set (18).

Information theory has previously been used to build two models to evaluate and predict Fur binding sites (12,19). In one case the model was built using some sites that had not been footprinted by Fur and were probably not aligned to maximize the information content (19). Both models used *ad hoc* variations of information theory to assign scores to the predicted binding sites, rather than classical information content in bits.

The most rigorous approach to model building is to create a data set comprised of only footprinted binding sites from one species. By restricting the data set to experimentally proven sites, one is certain that the model will reflect the binding characteristics of the protein; the use of a single species ensures that the protein and DNA binding sequences evolved together and therefore correspond to one another (20). Biases from previous models are thereby avoided. The resulting experimentally supported model is then scanned across the entire genome of the species, looking for sequences that contain a positive amount of

information as evaluated by the weight matrix (18). Sequence walkers, which are graphical representations of individual binding sites, then display probable binding sites on the genome based on the model of footprinted sites (21). This method was successfully used to discover that the OxyR transcription factor controls the expression of the *fur* gene (22), to identify additional sites for proteins such as Fis, SoxS, and OxyR (23–25), and to characterize splice junctions (26). In this study, information theory has allowed us to identify new Fur binding sites, thirteen of which we confirmed experimentally.

Materials and Methods

Programs

Programs used in this study are available at http://www.lecb.ncifcrf.gov/~toms/. A web-based tool for searching for Fur sites is available at
http://www.lecb.ncifcrf.gov/~toms/delilaserver.html.

Creating the Fur Model

Twelve experimentally confirmed, footprinted sequences from *E. coli* were extracted from the *E. coli* genome by the **delila** program (Fig. 1) (27). The model was comprised of sites from the promoters of the genes *cir*, *fecA, fecIR, fur, sodA, iucA, tonB,* and *hlyCABD*, along with two bidirectional promoter regions for the genes *fepA-fes* and *fepB-entC* (28–36). The promoter *fepB-entC* has two distinctly protected regions; each region was included in the data set as individual sequences (*fepB* and *entC*). The promoter of *iucA* has an exceptionally long secondary footprint (7) and so two regions were used (*iucA1* and *iucA2*). The complement of each footprint was also included, since Fur binds as a dimer (6, 7). The program **malign** was used to obtain an alignment of the sequences (Fig. 1) that maximizes the information content of the model (16). The range of the final model was from -12 to $+12$ base pairs, chosen from the concentration of significant sequence conservation observed in the logo (Fig. 1).

The validity of the model was tested by scanning it across the promoter regions of the genes used to create it, using the programs **scan** and **lister** to create sequence walkers (18, 21). The model is verified if the walkers correspond to DNase I footprint regions. For further verification, the model was also scanned across synthetic Fur binding sites containing GATAAT repeats in oligonucleotides that have been previously footprinted (Fig. 5) (5).

Scanning for Fur Binding Sites

The second law of thermodynamics sets a theoretical lower bound for the information content of an individual binding site (R_i) at zero bits (18) but several other cutoff levels were used for various purposes. These cutoffs were at -200 bits for forced walkers, 3 bits for a likely natural cutoff, and 16 bits for genomic scanning, as described below.

In addition to the footprinted promoters, several other promoters of genes proposed to be iron-regulated were scanned: *exbB*, *exbD*, *feo*, *fldA*, *yhhX*, *gpmA*, *ygaC*, and *nohA* (9, 37, 38). If no sequence walkers appeared in the scan, walkers were 'forced' to appear by lowering the allowed individual information content for a site to -200 bits, which causes the model to detect all positions. The highest information content of all possible sequences in the region

could then easily be identified; if that value was less than zero, then the probability of a sequence existing in the region that would fit the model is small, which implies that the protein should not bind (18).

The **scan** program parameters were normally set to record all sites with individual information content greater than 3 bits. No footprinted sites had an information content less than 6.7 bits, and a value near 3 bits may be a fundamental bound for true binding sites (26).

For the whole genome scan, all sites with R_i values greater than 16.0 bits were extracted. This value was chosen simply to allow for a manageable set of regions for further analysis. Groups of walkers that were within 200 bases of each other were identified using the **localbest** program, and the strongest one was selected to represent the region. This ensured that each region in the revised data set was unique. Sites that had been included in the model were also removed from the data set. With the remaining sites, a new set of **delila** instructions was created to extract 400-base regions surrounding the selected sites. The set of sequences produced by **delila** contained the strongest sites from the genome scan; this was then scanned with the same model as originally used to find the sites in the genome. The subsequent **lister** map displayed all the sites in the new set of sequences that fit the model.

To further focus our set of potential Fur sites, we searched the genome for strong Fur sites that overlapped exceptionally strong promoters within 200 bases of a gene start. To quantify the strength of the promoters, we used an information theory based flexible σ^{70} binding model that uniformly takes into account the information present in the -10, the -35, and the uncertainty of the spacing between them (Shultzaberger and Schneider, in preparation). ■ **switch to Shultzaberger.Schneider-flexprom2003 if possible** We have successfully used flexible modeling for prokaryotic ribosome binding sites (39). We then further narrowed this set of sites based on the function of the potentially repressed genes, leaving us with 7 predicted Fur repressed promoters for testing: *yoeA*, *fepD-entS* [formerly *ybdA*], *gpmA*, *yhhX-yhhY*, *fhuA*, *nohA*, and *oppA*.

■ **I put a temporary page break to make viewing the next section easy.**

Footprinting

■ **The *fhuF* promoter construct pGSO129 (25) was used to test for Fur binding. Purified Fur protein, generously provided by C. Outten and T. O'Halloran, was incubated with a Mn^{2+}-containing buffer according to de Lorenzo *et al.* (7). DNase I footprinting then was carried out as described previously (40). The 240 bp *Bam*HI-*Eco*RI fragment of pGSO129 was labeled with [γ-^{32}P]ATP at either the *Bam*HI site (top strand) or the *Eco*RI site (bottom strand). The labeled fragments were incubated with 0, 50, 100, 200, 400 mM purified Fur protein at room temperature for 5 min. The samples then were subjected to limited DNase I digestion, purified and separated on 8% polyacrylamide sequencing gels. →**

(new version suggested by Gigi. Essentially only the first sentence was changed:)
To generate the *fhuF* promoter construct (pGSO129) used to test for Fur binding, a 240 bp fragment amplified by PCR from genomic DNA (using the primers 5′-GCG GCT GGA GAT GAA TTC GCC AGA TG and 5′-GCC CTG CAA TCA GGG ATC CCG GCA GC) was cloned into the BamHI and EcoRI sites of pUC18 (25). Purified Fur protein, generously provided by C. Outten and T. O'Halloran, was incubated with a Mn^{2+}-containing buffer according to de Lorenzo *et al.* (7). DNase I footprinting then was carried out as described previously (40). The 240 bp BamHI-EcoRI fragment of pGSO129 was labeled with [γ-^{32}P]ATP at either the BamHI site (top strand) or the EcoRI site (bottom strand). The labeled fragments were incubated with with 0, 50, 100, 200, 400 mM purified Fur protein at room temperature for 5 min. The samples then were subjected to limited DNase I digestion, purified and separated on 8% polyacrylamide sequencing gels.

Gel Mobility Shift Assays

Two sets of oligonucleotides containing known and predicted Fur binding sites in *E. coli* were designed and synthesized (Oligos Etc). All oligonucleotides were self-complementary, had a 5′ overhang, and contained a hairpin loop. Oligos *exbB*, *exbD*. and *fhuF* contained a hairpin of the sequence 5′-GCGAAGC-3′, while the other twelve oligos (*yoeA*, *fepD-entS* [formerly *ybdA*], *gpmA*, *yhhX-yhhY*, *fhuA*, *nohA*, *oppA*, *gspC*, *yhaU*, *yahA*, *fadD*, and *ygaC*) contained a hairpin of the sequence 5′-ACGATCGC GCGAAGC GCGATCGT-3′ in the center. Such loops form a structure which is stabilized by base pairing between G_3 and A_5 of the central seven bases of each loop (41), and it is convenient for use in DNA mobility shift assays because of the exact equimolar concentrations of the complementary strands and its high stability (23).

Three oligos containing the promoter regions for *exbB* and *ygaC* and the upstream region of *exbD* were created to test previously published consensus sequence predictions (Fig. 8, ref. 37, 38). Seven oligos contained potential Fur controlled promoters identified using both the Fur and a σ^{70} model as described above (*yoeA*, *fepD-entS* [formerly *ybdA*], *gpmA*, *yhhX-*

yhhY, *fhuA*, *nohA*, and *oppA*). We were also interested in whether Fur bound intragenically, so four oligos were synthesized that contained strong predicted sites found within gene coding regions (*gspC*, *yhaU*, *yahA*, *fadD*). As a positive control, an oligo was created containing the primary *fhuF* binding site.

One mobility shift assay was performed using only the oligos *exbB*, *exbD*, and *fhuF* (Fig. 9). The oligonucleotides were labeled by a fill-in reaction with Taq DNA polymerase. The 50 μl reaction mixture (1 μM oligo, 2 mM $MgCl_2$, 1x PCR buffer, 25 units Taq polymerase (GibcoBRL®), and 10 μM tetramethylrhodamine-6-dUTP (NEN)) was incubated at 72°C for 30 minutes, followed by two phenol/chloroform extractions. The DNA was then purified with a 10% polyacrylamide gel in TBE buffer.

The labeled oligonucleotides were then incubated in Fur binding buffer (10 mM Bis-Tris-HCl pH 7.5, 5 μg/ml sonicated salmon sperm DNA, 5% glycerol, 100 μM $MnCl_2$, 100 μg/ml BSA, 1 mM $MgCl_2$, 40 mM KCl) with Fur protein at various concentrations at 37°C for 13 minutes (7). Samples were electrophoresed on a 5% polyacrylamide gel in Fur electrophoresis buffer (0.1 M Bis-Tris-HCl pH 7.5, 10 mM $MnCl_2$) at 120V for about 2 hours.

Bands were visualized with an FMBIO II fluorescent scanner (Hitachi), with an excitation wavelength of 532 nm and a 585 nm filter for detection of tetramethylrhodamine.

■ THIS PARAGRAPH IS BEING CHECKED AND REVISED BY ZEHUA: A second mobility shift assay was then performed using all fifteen oligos, using procedures similar to the first assay (Fig. 10). ■ Zehua: check the molarity of all components and amount of Taq polymerase in reaction mixture! The oligonucleotides were diluted 1:5 ■ Ilya asks Zehua: Was it in TE or water or something else? and boiled for 10 minutes, then placed on ice to prevent dimerization and trimerization. The 20 μl of reaction mixture (10 pmol oligo, 2 mM $MgCl_2$, 1x PCR buffer, 25 units Taq polymerase (GibcoBRL®), and 10 μM tetramethylrhodamine-6-dUTP (NEN)) was incubated at 72°C for 1 hour. 80 μl of ddH_2O was added to the mixture, followed by two phenol extractions, a phenol/chloroform extraction and a chloroform extraction. The oligos were diluted 1:20 for incubation in Fur binding buffer, followed by gel electrophoresis and band visualization as described above.

→

■ CURRENT VERSION: A second mobility shift assay was then performed using all fifteen oligos, using procedures similar to the first assay (Fig. 10). The 20 μl of reaction mixture ■ *WHAT REACTION???* (20 pmol oligo, 2 mM $MgCl_2$, 1x PCR buffer, 5 units Taq polymerase (GibcoBRL®), and 10 μM tetramethylrhodamine-6-dUTP (NEN)) was incubated at 72°C for 1 hour. 80 μl of ddH_2O was added to the mixture, followed by two phenol extractions, a phenol/chloroform extraction and a chloroform extraction. The labeled oligonucleotides were diluted 1:5 ■ *IN WHAT??? WATER?* and boiled for 10 minutes, then placed on ice to prevent dimerization and trimerization. 5 μl of oligos were used for incubation in Fur binding buffer, followed by gel electrophoresis and band visualization as described above.

Results

Fur Binding Model

Sequence conservation in twelve footprinted Fur binding sites was evaluated by creating an information theory model (15). Because Fur binds as a dimer (42), the footprinted Fur sequences and their complements were used to create the model (Fig. 1). ⇐Fig 1

The sequences were first aligned using the program **malign**, which maximizes the information content of sequences by shuffling them back and forth (16). This method of multiple alignment provides not only the best (highest information content) alignment but also variant alignments that are slightly worse. By repeating the alignment process starting from the best alignment, we obtained alternative alignments relative to the best alignment. An alignment of n sequences is expressed as a set of n numbers, each of which represents the number of base pairs that the sequence has been displaced. Thus when we repeat the alignment, the best alignment is obtained again and since it has no displacements, it is a vector consisting of all zeros (Fig. 2A, Alignment 1). Other sequence alignments have lower information content ⇐Fig 2 and occur less frequently. The components of these secondary vectors often add up to six, indicating that the best alignment was shifted so that two bases that were initially six bases apart become lined up. This is a strong indication that Fur binds in clusters with 6-base separation.

The alignment that has the strongest information content occurred most frequently (Fig. 2B), and produced a sequence logo (17) (Fig. 1) containing a strongly conserved region between coordinates -12 and $+12$, with $R_{sequence}= 19.5 \pm 1.5$ bits (15). The sequence logo follows a sine wave with a wavelength of 10.6 bases, suggesting that Fur binds to one face of the DNA (27,43,44). From the logo, major and minor groove contacts can be predicted (27). Positions ± 4, ± 5, and ± 6 exceed 1 bit and so are likely to represent major groove contacts (27). Positions ± 2 and ± 3 appear to be major groove N7 contacts, since both adenine and guanine base pairs are found at -2 and -3, and their complementary bases are found at $+2$ and $+3$. Position 0 approaches the maximum information for a minor groove contact of one bit; Fur may exclude the N2 amino group of guanine at this position in the minor groove when it binds to the DNA (45). Positions ± 1 would appear to be major groove contacts because they do not contain equiprobable A and T, but they are so far into the minor groove that the conservation is more likely to be caused by overlapping sites (see below) (27).

Scans of Published Footprints

All twelve footprinted sequences displayed clusters consisting of multiple overlapping Fur walkers (Table 1); in these clusters, the majority of the walkers were separated by six bases, ⇐Table 1 an example of which is shown in Fig. 3.

Fur has been documented to exhibit 'secondary footprinting', protecting extended regions ⇐Fig 3 of DNA under higher protein concentrations (Fig. 4) (29,32,35,46). The strongest walker in ⇐Fig 4 each of these regions covers only part of the footprint; on average these walkers account for only 60% of the protected regions, and so they do not account for the footprints. However, the entire region of protein protection is adequately accounted for by several lower-information content walkers that appear in clusters, since walkers with $R_i > 0$ bits cover $125 \pm 12\%$ of the protein-protected region (Table 1). This mean is above 100% because several of the clusters contain low-information content walkers that extend past the range of the footprints, creating more than 100% coverage of the footprint by the walkers.

In a previously published study, Escolar *et al.* synthesized oligonucleotides that contained

repeats of the sequence GATAAT and determined Fur binding by footprint experiments (5). No footprint was seen with one insert; correspondingly, no sequence walkers were observed in that sequence (Fig. 5). The 2-insert sequence had a weak interaction with Fur at high protein concentrations and a 3.6 bit walker appeared. The majority of the information content of this walker is contained in the part that overlaps the protected insert sequence. Sequence walkers accurately matched the footprints of longer synthetic binding sites. The parts of walkers extending into fainter protection exhibit lower information content. These results show that the Fur model accurately predicts binding to both natural and synthetic sequences. ⇐Fig 5

Whole Genome Scan

6837 sites were found in the scan of the *E. coli* genome with a 3-bit cutoff, which is a likely biologically important lower bound (26). These results are available at http://www.lecb.ncifcrf.gov/~toms/papers/fur/. Out of these, forty novel regions were found that contained at least one predicted Fur binding site with an information content over 16.0 bits (an arbitrary cutoff to locate significant regions) (Table 2). Clusters of walkers were present in all 40 regions ($R_i > 0$). As observed in the footprinted regions, the walkers in the clusters were often separated by six bases. In addition to the characteristic 6-base spacing, some walker clusters also displayed 3-base spacing. ⇐Table 2

Of the 40 strongest sites, 10 were inside genes (47). Since there are so many predicted sites that are probably in promoter regions for genes, Fur appears to be a pleiotropic gene regulator. The strongest sequence walker in the entire genome, at 26.2 bits, was found in the promoter region of the *fhuF* gene, formerly known as *yjjS* (48).

Fur Footprints at the *fhuF* promoter

Sequence walkers predict that the strong *fhuF* site is surrounded by other Fur sites, and these are within one of two distinct clusters of walkers which are clearly separated by a 24 base-pair gap (Fig. 6). DNase I footprinting (Fig. 7) shows that there are indeed two Fur-protected regions in the *fhuF* promoter. The regions protected from DNase I digestion fit the sequence walkers very well, with only a slight overhang at the 5′ ends. The region downstream is less protected at lower concentrations of Fur. ⇐Fig 6 ⇐Fig 7

Scans of Other Proposed Fur-regulated Genes

Many genes have been proposed to be Fur-regulated by comparing consensus sequences to promoter regions and also by homology to systems in other organisms. Kammler *et al.* proposed 'Fur boxes' in the promoter region of *feo* using a consensus sequence (9). Six sequence walkers were found in this region, in close proximity to but not exactly matching the consensus sequences marked by the authors (1.4, 9.5, 0.6, 12.4, 15.5, and 2.8 bits at 3537601, 3537617, 3537664, 3537670, 3537676, and 3537682, respectively). The same authors have confirmed that Fur does bind to this region *in vivo*, but the exact Fur binding site has not been determined by footprint experiments.

In *Klebsiella pneumoniae* and *Anacystis nidulans*, the gene encoding flavodoxin (*fldA*) has been observed to be regulated by the respective Fur homologue (49). When the *E. coli fldA* promoter region was scanned, three weak walkers were found (3.3, 1.0, and 1.9 bits at 710683, 710751 and 711263, respectively), indicating that perhaps Fur also regulates

flavodoxin in *E. coli*.

Vassinova and Kozyrev used an *in vivo* selection to locate Fur sites on *Sau*3A fragments from the *E. coli* genome (38). The five regions from Figure 2 of their paper were analyzed using sequence walkers. Using an unidentified consensus sequence, *yhhX* (3578883 to 3578642) was predicted by Vassinova and Kozyrev to have two Fur binding sites. We found two walkers having 16.7 and 13.5 bits at 3578665 and 3578659. An additional site of 8.0 bits is found upstream of the two stronger sites, at 3578712. The promoter region of *gpmA* (786731 to 787328, identified as *pgm* in reference 38), was predicted by consensus to contain two sites. Five walkers were found, consisting of 1.2, 17.5, 12.3, 20.0, and 1.1 bits at 786844, 786850, 786853, 786856, and 786893, respectively. No walkers were found at the consensus sequence-predicted site in front of *ygaC* (2798846 to 2798161). However, the *Sau*3A fragment contains 5 walkers having 6.4, 3.0, 7.3, 12.6 and 7.3 bits at 2798762, 2798752, 2798746, 2798740, and 2798509, respectively. For *nohA* (1634711 to 1634593), as predicted by consensus, four walkers were found having 0.3, 21.4, 3.1 and 6.3 bits at 1634633, 1634627, 1634624, and 1634621, respectively. The fifth region was *fhuF* (4603463 to 4603166). Figures 6 and 7 show the predicted walkers and our footprints in this region. Only the high affinity region was identified by the consensus sequence. Thus four of the *Sau*3A fragments selected to have Fur binding sites *in vivo* were identified in our genome scan (*yhhX*, *gpmA*, *nohA* and *fhuF*, Table 2). The fifth *Sau*3A fragment (*ygaC*) was not included in our set of predicted genomic binding sites simply because it did not have any sequence walkers above our 16 bit cutoff. Our model does predict Fur binding up to 12.6 bits in the *ygaC* promoter region at positions different from the site predicted by a consensus sequence (38).

A microarray analysis by McHugh *et al.* found 143 genes to be either directly or indirectly affected by Fur (12). We scanned the promoter regions of these genes with both the Fur and promoter models, using a lower bound cutoff of 6.7 bits for Fur, the strength of the weakest site in our model, and searched 300 bases upstream to 30 bases downstream of the translational start. McHugh *et al.* used an altered information theory approach for modeling Fur binding (50) to identify which genes are under direct Fur control. Our model identified sites in all genes that their model did, as well as 17 others. Five of these, *fhuE*, *ftnA*, *exbB*, *fepE*, and *pqqL*, were repressed according to the microarray data, and we found a strong overlapping σ^{70} site, suggesting direct repression of these genes by Fur. Two cases of direct activation by Fur were suggested by the microarray data and our scan. In *yfaH*, a 7.8 bit Fur site is predicted to be immediately upstream of the -35, while in *narG*, a 7.0 bit site is 91 bases upstream but a 9.1 bit Fis site between them could bend the DNA to bring bound Fur to the promoter (23). Although direct activation by Fur has not been shown, the positions of these sites relative to the promoter are comparable to those of direct activators. The remaining 10 sites, *narK*, *frdA*, *ymfE*, *ydfK*, *cspI*, *sodB*, *ybiJ*, *sufC*, *feoB*, and *nirC* do not have an obvious mechanism of control.

Eick-Helmerich and Braun matched a Fur consensus sequence to the promoters of *exbB* and *exbD* (37). Three walkers were found in the *exbB* promoter region, completely covering the 'Fur box' (Fig. 8). The *exbD* promoter region showed no walkers with $R_i > 0$, even ⇐Fig 8
though it had also been predicted to contain a Fur binding site using the same consensus sequence as used in the *exbB* promoter. To ensure that the absence of walkers was not due to two or three strongly negative bases, the forced walker method was used on the *exbD* promoter, as described in Materials and Methods. In the promoter region for *exbD*, no site

appeared with an information content greater than −8.8 bits between coordinates 3149340 and 3149540; around the designated 'Fur box', as shown in Figure 8, only one walker of −10.6 bits was found, indicating that Fur should not bind.

Fur Binds to *exbB* but not *exbD*

The ability of Fur to bind to the *exbB* and *exbD* promoters was examined by a gel mobility shift assay using the *fhuF* promoter Fur site as a positive control. The potential binding sites from each promoter region were incorporated into double-stranded hairpin oligonucleotides (Fig. 8). As predicted by sequence walkers, the *exbB* and *fhuF* oligos were shifted by Fur whereas the *exbD* only shifted non-specifically at extremely high concentrations of Fur (Fig. 9 and 1.3 μM, data not shown). The *fhuF* sequence had additional shifts as the concentration of Fur increased. Under higher concentrations of Fur, *exbB* also displays multiple shifts (1.3 μM, data not shown). These additional shifts are consistent with the prediction of multiple strong Fur sites (Fig. 8). McHugh *et al.* found that both *exbB* and *exbD* were regulated by both iron and Fur, but their probability model did not detect a binding site in the *exbB* promoter while our information theory model successfully predicted that Fur binding site. ⇐Fig 9

Additional Fur Binding Sites

In a second round of experiments, all oligos containing a predicted Fur binding site (see Methods) shifted when incubated with Fur protein (Fig. 10). The *exbD* oligo did not shift while *ygaC* only showed weak binding. ⇐Fig 10

Discussion

In our study we used footprinted *E. coli* sequences to create an information theory model of Fur binding. The model appears to approximate the binding characteristics of Fur more fully than models used in previous studies, which depended on consensus sequences, data from multiple species, and sequences which were not footprinted (1, 28, 38). The rigorous approach revealed new binding sites, disproved two sites predicted by a consensus sequence, and clarified the manner in which the protein binds.

Our genome scan identified two clusters of sequence walkers in the promoter region of *fhuF* (Fig. 6). The upstream cluster was stronger and contained the highest information content walker in the entire genome (26.2 bits). Fur binding has been previously established in this region through genetic methods, and the *fhuF* promoter was utilized in a 'Fur titration assay' (1). This region was also found in an *in vivo* selection, but only one of the clusters was located by the consensus sequence method (38). Our DNase I footprints (Fig. 7) revealed two Fur binding regions of differing affinities, with a strong correlation between the footprinted regions and the regions covered by sequence walkers (Fig. 6). The high-affinity region corresponds almost exactly to the strong upstream cluster of walkers containing the 26.2 bit walker. The low-affinity region roughly matches the downstream cluster of walkers, but with less precision than the high-affinity region. The promoter region of *fhuF* is apparently shared with that of the open reading frame *yjjZ*. Not only is the entire intergenic region involved in Fur binding, but there are also two distinct OxyR binding sites corresponding to each of the Fur clusters (25).

In addition to correctly identifying the binding range in strong information content re-

gions, our model is also capable of discerning between binding sites and non-binding sites. Previous authors used the consensus sequence method to identify 'Fur boxes' in the promoter regions of *exbB* and *ygaC* and upstream of *exbD* (37, 38). While three sequence walkers appeared in the *exbB* promoter, none appeared in the promoter of *exbD* (Fig. 8). With the highest information content of any of the forced walkers in the *exbD* promoter region being -8.8 bits, the entire sequence is not compatible with the information theory model of the natural sites. The probability that a -8.8 bit site is part of the natural population is 1×10^{-10}, which indicates that Fur would not bind to this region (18). Gel mobility shift assays with synthetic oligonucleotides confirmed that the consensus-predicted site of *exbD* does not bind the Fur protein, while the promoter region of *exbB* does (Fig. 9).

Additional gel shifts confirmed predictions of Fur binding sites made by the information theory model in both promoter and intergenic regions (Fig. 10). All oligos that contained sequence walkers with information content greater than zero exhibit Fur binding. The set of oligos included the five regions predicted by Vassinova and Kozyrev using a consensus sequence (*yhhX*, *gpmA*, *nohA*, *fhuF* and *ygaC*; ref. 38). A faint shift can be observed with the *ygaC* oligo under high concentrations of Fur. The strongest sequence walker on that oligo was -2.2 bits, and contains a G at position 0 that is not seen in the model. If Fur accepts a G at this position, then when this site is included in a revised model, the information content of the walker will increase, which allows for the possibility of binding Fur at high concentrations. The mobility shifts also confirmed a predicted binding site in the *fepD-entS* [formerly *ybdA*] bidirectional promoter region, which was recently footprinted by others (51). While the banding patterns of the shifted oligos may hold clues to the number of Fur dimers that are bound, further work is needed to exactly correlate the amount of Fur bound and the number of walkers that appear on each oligo.

Recent publications have described Fur binding to sequences containing mutations in the *fepD-entS* [formerly *ybdA*] promoter region (51) and synthetic oligonucleotides composed of variations of a Fur consensus sequence (5, 52), as well as a microarray analysis of Fur-regulated genes (12). Sequence walkers found by our information theory model are consistent with the footprints, gel shifts, and expression levels obtained in these studies. Specifically, the walkers are able to explain the results of the Fur binding studies more fully than the consensus sequence model (*e.g.* Fig. 5). Seventeen of the Fur-regulated genes found by the microarray study (12) were predicted by sequence walkers to have a Fur binding site that was not detected by the method used by McHugh *et al.* A sequence logo of *B. subtilis* Fur binding sites (53, 54) is significantly different from the *E. coli* model (Fig. 1). We built an individual information weight model starting from the sequences provided by Baichoo *et al.* by including the complementary sequences, removing unproven sites, and aligning to maximize information content (data not shown). *B. subtilis* sequence walkers had significantly different information contents when compared to those created by the *E. coli* model and they also generally correspond to the gel shift patterns better than the *E. coli* model (data not shown), indicating that the Fur proteins in the two organisms differ in their recognition of binding sequences and that cross-species analysis of Fur binding should be avoided. Interestingly, the *E. coli* logo (Fig. 1) is closely related to that of *B. subtilis* (53) (Figure 4b): positions -6 to -4 of *E. coli* correspond to positions 13 to 15 of *B. subtilis* and positions -3 to -1 of *E. coli* correspond to positions 10 to 12 of *B. subtilis*. In other words, two pieces of the monomer appear to have interchanged during evolution.

Several lines of evidence indicate that the Fur dimer binds in clusters, with six-base spacing between the individual dimers. First, in the sequence logo derived from the footprinted natural binding sites (Fig. 1), positions -12 to -10 resemble positions -6 to -4 and 0 to $+2$. From the symmetry of the model, the positions -2 to 0, $+4$ to $+6$, and $+10$ to $+12$ are also similar. These similar parts are spaced six bases apart. Second, multiple alignments show that variations from the best alignment are often found by shifting sequences by six bases (Fig. 2). Third, six-base spacing is observed in the scans of footprinted binding sites (Figs. 3, 4, 5, 6; Tables 1 and 2). The sequence walkers appear in clusters, with each individual walker most often spaced six bases apart from its neighbors. These clusters fit extraordinarily well to primary and secondary footprints in scans of both natural and synthetic binding sites. Six-base spacing has been noted by other workers (5, 38), but the use of sequence walkers revealed extensive clusters containing Fur sites spaced apart by 6 and 3 bases. This phenomenon was not detected by a previous application of information theory to Fur binding sites (19). Since the dimeric motif is reiterated every 6 bases (Fig. 2), our model is most compatible with dimeric Fur binding to overlapping sites that are spaced apart by 6 bases. Thus, although Fur binds as a dimer to sites with a two-fold rotational axis of symmetry, the sites also have translational symmetry.

Many difficulties in understanding Fur binding sites can be attributed to the choice of consensus sequences as a model. In contrast to sequence walkers, the consensus sequence method ignores the varying importance of bases by treating mismatches equivalently. In addition, the consensus method does not have a criterion for an acceptable number of mismatches, whereas the natural cutoff for sequence walkers is at zero bits (55). The success of the information theory model suggests that global predictions of the number of Fur sites in the genome may be reasonable. 92 walkers appeared in the footprinted regions (Table 1). We also predicted that there are at least 459 additional sites in the *E. coli* genome (Table 2), giving a total of 551 sites in clusters containing a strong site. However, in a full scan of the genome we found 6837 sites over 3.0 bits. In order to choose this many or more sites out of all the possible sites in the genome (4.7×10^6 bp), Fur needs less than $R_{frequency} = \log_2(\frac{4.7\times10^6}{6837}) = 9.4$ bits of information (15). The model has an $R_{sequence}$ of 19.5 bits in the strongly conserved region from -12 to $+12$; this is in excess of the amount of information needed to bind one Fur molecule by 10.1 bits. However, the information content of all sites ($R_i > 0$) in the clusters specified in Table 1 is 10.8 ± 0.4 bits. So the 19.5 bits represents a maximum and the 10.8 bits a minimum, suggesting that Fur sites average between 10 and 19 bits. It has been observed that the amount of information in binding sites ($R_{sequence}$) is close to the amount of information needed to find the binding sites ($R_{frequency}$). Depending on the number of binding sites in the genome and which weak sites in clusters are actually bound, $R_{sequence}$ may be close to $R_{frequency}$ for Fur (15, 20).

The excess information, as well as the self-similarity of the sequence logo, suggests that more than one Fur molecule binds to the same stretch of DNA (15, 56). Using electron microscopy, high-order multimers of Fur binding to the DNA have been observed, with Fur appearing to wrap around the DNA helix at successive sites (2). Proteins sharing the DNA by binding to it at the same time, as implied by the helix-wrapping hypothesis, can be easily inferred from the clusters of walkers spaced apart by six bases. It is possible that the Fur dimer first binds to the sites represented by the strongest walkers (in bits); once all of the 'primary' sites are bound by Fur, the remaining dimers will bind to the next strongest sites,

which overlap the strongest sites. This hypothesis is supported by the scans of the *iucA* (Fig. 4) and *sodA* (data not shown) promoter regions, which both show longer footprints with increased protein concentration and weaker sequence walkers in the extended region. The final DNA-Fur complex would then have proteins bound to the DNA at sites that are six bases apart, and since B–form DNA has one turn every 10.6 bases (57, 58), the proteins would be roughly on opposite faces of the molecule, overlapping each other. This overlapping between binding sites could create the moderately conserved regions observed on the flanks of the sequence logo, since two proteins would be reading information from the same DNA sequence. The overlap may be involved in cooperativity of Fur binding (59, p.864).

Baichoo and Helmann recognized a so-called 7-1-7 dimer binding mode of Fur (52), and Lavrrar and McIntosh have also proposed a model of Fur dimers binding at overlapping sites on opposite faces of the DNA (60). This is similar to our model in that we found that a dimeric model could account for the observed footprint data. Baichoo and Helmann observed that two dimers spaced 6 bases apart could bind to opposite faces of the DNA and demonstrated that the observed molecular weight of the slow mobility band is consistent with the presence of two Fur dimers. Our information-theory based model predicts this mode, but in addition it predicts two other distinct modes: multiple sets of Fur dimers at various spacings and Fur spacings 3 bases apart. Consistent with the findings of these two groups, our gel shifts support the idea that two Fur dimer molecules at 6 base spacing can bind simultaneously. 3 base spacing may represent a self-competition mode (61).

Based on our current understanding of Fur binding, we now suggest that a second, more realistic model could be built by trimming the initial dimeric -12 to $+12$ model (Fig. 1) to a smaller range of -7 to $+7$. To be consistent with footprint data, presumably the protein would still protect a region from -12 to $+12$. Such a model would be consistent with several observations. First, the model allows the protein to pack alternatively on top and bottom 'faces' of the DNA every six bases apart without collisions. Second, the flanks in our current model would be a consequence of the clustered nature of Fur binding, representing frequent binding of a neighboring dimer. Because Fur sites come in clusters, removing the outer flanks does not affect the model much (data not shown). Third, the reduced range would have an information content of 14.8 bits for a ± 7 model, which is closer to the information needed to find 6837 sites in the genome (9.4 bits) and which could account for the excess information in the ± 12 model. However, because the sites overlap, even with a reduced range of ± 7 the model would still represent binding from both sides of the DNA. It is not clear how to deconvolve the logo into one that represents binding from only one face of the DNA. It may be possible to perform a SELEX experiment that selected for a single binding protein, but SELEX can generate artifacts (62), and there may be no way to determine if this were the case.

The proteins coded by the genes that are predicted to be regulated by Fur (Table 2) encompass a variety of functions in the cell, including DNA metabolism (DNA repair and replication) and enzymology (biosynthesis and general cell metabolism), as well as the well-established regulation of iron-transport systems. It has been determined that the Fur protein is very abundant in *E. coli* cells, numbering around 5000 molecules per cell (22), close to the number of sites we found in the entire genome (6837 sites >3.0 bits). Thus Fur is much more abundant than other regulators, which average 100 molecules in each cell (63). The high levels suggest that Fur is a pleiotropic regulator, much like the protein Fis, which is

present in up to 50,000 dimers per cell (23). Fur appears to be an effective global regulator, numerous and capable of regulating many vital systems in the *E. coli* cell.

Acknowledgments

We thank Tom O'Halloran and Caryn Outten for Fur protein; Pete Rogan for helping to develop the local best concept; Elaine Bucheimer for writing the **localbest** program; and Lakshmanan Iyer, Brent Jewett, Danielle Needle, Xiao Ma, Shu Ouyang, Denise Rubens, and Bruce Shapiro for their helpful comments and discussions. K.A.L. also thanks the National Cancer Institute at Frederick for sponsoring the Werner H. Kirsten Student Intern Program. This publication has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract #NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES

1. Stojiljkovic, I., Baumler, A. J., and Hantke, K. (1994) Fur regulon in gram-negative bacteria. Identification and characterization of new iron-regulated *Escherichia coli* genes by a *fur* titration assay [published erratum appears in *J. Mol. Biol.* 1994 Jul 15;240(3):271]. *J. Mol. Biol.,* **236**, 531–545.

2. Le Cam, E., Frechon, D., Barray, M., Fourcade, A., and Delain, E. (1994) Observation of binding and polymerization of Fur repressor onto operator-containing DNA with electron and atomic force microscopes. *Proc. Natl. Acad. Sci. USA,* **91**, 11816–11820.

3. Calderwood, S. B. and Mekalanos, J. J. (1987) Iron regulation of Shiga-like toxin expression in Escherichia coli is mediated by the fur locus. *J. Bacteriol.,* **169**, 4759–4764.

4. Niederhoffer, E. C., Naranjo, C. M., Bradley, K. L., and Fee, J. A. (1990) Control of Escherichia coli superoxide dismutase (sodA and sodB) genes by the ferric uptake regulation (fur) locus. *J. Bacteriol.,* **172**, 1930–1938.

5. Escolar, L., Perez-Martin, J., and de Lorenzo, V. (1998) Binding of the Fur (Ferric Uptake Regulator) repressor of *Escherichia coli* to arrays of the GATAAT sequence. *J. Mol. Biol.,* **283**, 537–547.

6. Bagg, A. and Neilands, J. B. (1987) Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli. *Biochemistry,* **26**, 5471–5477.

7. de Lorenzo, V., Wee, S., Herrero, M., and Neilands, J. B. (1987) Operator sequences of the aerobactin operon of plasmid ColV-K30 binding the ferric uptake regulation (*fur*) repressor. *J. Bacteriol.,* **169**, 2624–2630.

8. Althaus, E. W., Outten, C. E., Olson, K. E., Cao, H., and O'Halloran, T. V. (1999) The ferric uptake regulation (Fur) repressor is a zinc metalloprotein. *Biochemistry,* **38**, 6559–6569.

9. Kammler, M., Schon, C., and Hantke, K. (1993) Characterization of the ferrous iron uptake system of *Escherichia coli*. *J. Bacteriol.,* **175**, 6212–6219.

10. Desai, P. J., Angerer, A., and Genco, C. A. (1996) Analysis of Fur binding to operator sequences within the *Neisseria gonorrhoeae fbpA* promoter. *J. Bacteriol.,* **178**, 5020–5023.

11. Heidrich, C., Hantke, K., Bierbaum, G., and Sahl, H. G. (1996) Identification and analysis of a gene encoding a Fur-like protein of *Staphylococcus epidermidis*. *FEMS Microbiol. Lett.,* **140**, 253–259.

12. McHugh, J. P., Rodriguez-Quinones, F., Abdul-Tehrani, H., Svistunenko, D. A., Poole, R. K., Cooper, C. E., and Andrews, S. C. (2003) Global iron-dependent gene regulation in *Escherichia coli*. A new mechanism for iron homeostasis. *J Biol Chem,* **278**, 29478–29486.

13. Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Tech. J.,* **27**, 379–423, 623–656 http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

14. Pierce, J. R. (1980) An Introduction to Information Theory: Symbols, Signals and Noise, Dover Publications, Inc., New York second edition.

15. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.,* **188**, 415–431 http://www.lecb.ncifcrf.gov/~toms/paper/schneider1986/.

16. Schneider, T. D. and Mastronarde, D. (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics,* **71**, 259–268 http://www.lecb.ncifcrf.gov/~toms/paper/malign.

17. Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.,* **18**, 6097–6100 http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/.

18. Schneider, T. D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.,* **189**(4), 427–441 http://www.lecb.ncifcrf.gov/~toms/paper/ri/.

19. Panina, E. M., Mironov, A. A., and Gelfand, M. S. (2001) Comparative analysis of FUR regulons in gamma-proteobacteria. *Nucleic Acids Res,* **29**, 5195–5206.

20. Schneider, T. D. (2000) Evolution of biological information. *Nucleic Acids Res.,* **28**(14), 2794–2799 http://www.lecb.ncifcrf.gov/~toms/paper/ev/.

21. Schneider, T. D. (1997) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.,* **25**, 4408–4415 http://www.lecb.ncifcrf.gov/~toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.

22. Zheng, M., Doan, B., Schneider, T. D., and Storz, G. (1999) OxyR and SoxRS regulation of *fur*. *J. Bacteriol.,* **181**, 4639–4643 http://www.lecb.ncifcrf.gov/~toms/paper/oxyrfur/.

23. Hengen, P. N., Bartram, S. L., Stewart, L. E., and Schneider, T. D. (1997) Information analysis of Fis binding sites. *Nucleic Acids Res.,* **25**(24), 4994–5002 http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/.

24. Wood, T. I., Griffith, K. L., Fawcett, W. P., Jair, K.-W., Schneider, T. D., and Wolf, R. E. (1999) Interdependence of the position and orientation of SoxS binding sites in the transcriptional activation of the class I subset of *Escherichia coli* superoxide-inducible promoters. *Mol. Microbiol.,* **34**, 414–430.

25. Zheng, M., Wang, X., Doan, B., Lewis, K. A., Schneider, T. D., and Storz, G. (2001) Computation-Directed Identification of OxyR-DNA Binding Sites in *Escherichia coli*. *J. Bacteriol.,* **183**, 4571–4579.

26. Rogan, P. K., Faux, B. M., and Schneider, T. D. (1998) Information analysis of human splice site mutations. *Human Mutation,* **12**, 153–171 http://www.lecb.ncifcrf.gov/~toms/paper/rfs/.

27. Schneider, T. D. (1996) Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Meth. Enzym.,* **274**, 445–455 http://www.lecb.ncifcrf.gov/~toms/paper/oxyr/.

28. Griggs, D. W. and Konisky, J. (1989) Mechanism for iron-regulated transcription of the *Escherichia coli cir* gene: metal-dependent binding of Fur protein to the promoters. *J. Bacteriol.,* **171**, 1048–1052.

29. Angerer, A. and Braun, V. (1998) Iron regulates transcription of the *Escherichia coli* ferric citrate transport genes directly and through the transcription initiation proteins. *Arch. Microbiol.,* **169**, 483–490.

30. de Lorenzo, V., Herrero, M., Giovannini, F., and Neilands, J. B. (1988) Fur (ferric uptake regulation) protein and CAP (catabolite-activator protein) modulate transcription of *fur* gene in *Escherichia coli*. *Eur. J. Biochem.,* **173**, 537–546.

31. Hunt, M. D., Pettis, G. S., and McIntosh, M. A. (1994) Promoter and operator determinants for *fur*-mediated iron regulation in the bidirectional *fepA-fes* control region of the *Escherichia coli* enterobactin gene system. *J. Bacteriol.,* **176**, 3944–3955.

32. Tardat, B. and Touati, D. (1993) Iron and oxygen regulation of *Escherichia coli* MnSOD expression: competition between the global regulators Fur and ArcA for binding to DNA. *Mol. Microbiol.,* **9**, 53–63.

33. de Lorenzo, V., Giovannini, F., Herrero, M., and Neilands, J. B. (1988) Metal ion regulation of gene expression. Fur repressor-operator interaction at the promoter region of the aerobactin system of pColV- K30. *J. Mol. Biol.,* **203**, 875–884.

34. Young, G. M. and Postle, K. (1994) Repression of *tonB* transcription during anaerobic growth requires Fur binding at the promoter and a second factor binding upstream. *Mol. Microbiol.,* **11**, 943–954.

35. Frechon, D. and Le Cam, E. (1994) Fur (ferric uptake regulation) protein interaction with target DNA: comparison of gel retardation, footprinting and electron microscopy analyses. *Biochem. Biophys. Res. Commun.,* **201**, 346–355.

36. Brickman, T. J., Ozenberger, B. A., and McIntosh, M. A. (1990) Regulation of divergent transcription from the iron-responsive *fepB-entC* promoter-operator regions in *Escherichia coli*. *J. Mol. Biol.,* **212**, 669–682.

37. Eick-Helmerich, K. and Braun, V. (1989) Import of biopolymers into *Escherichia coli*: nucleotide sequences of the *exbB* and *exbD* genes are homologous to those of the *tolQ* and *tolR* genes, respectively. *J. Bacteriol.,* **171**, 5117–5126.

38. Vassinova, N. and Kozyrev, D. (2000) A method for direct cloning of Fur-regulated genes: identification of seven new Fur-regulated loci in *Escherichia coli*. *Microbiology,* **146**, 3171–3182.

39. Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E., and Schneider, T. D. (2001) Anatomy of *Escherichia coli* Ribosome Binding Sites. *J. Mol. Biol.,* **313**, 215–228 http://www.lecb.ncifcrf.gov/~toms/paper/flexrbs/.

40. Toledano, M. B., Kullik, I., Trinh, F., Baird, P. T., Schneider, T. D., and Storz, G. (1994) Redox-dependent shift of OxyR-DNA contacts along an extended DNA binding site: A mechanism for differential promoter selection. *Cell,* **78**, 897–909.

41. Hirao, I., Kawai, G., Yoshizawa, S., Nishimura, Y., Ishido, Y., Watanabe, K., and Miura, K. (1994) Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nucleases and heat. *Nucleic Acids Res.,* **22**, 576–582.

42. Bagg, A. and Neilands, J. B. (1987) Molecular mechanism of regulation of siderophore-mediated iron assimilation. *Microbiol. Rev.,* **51**, 509–518.

43. Papp, P. P., Chattoraj, D. K., and Schneider, T. D. (1993) Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.,* **233**, 219–230.

44. Schneider, T. D. (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucl. Acid Res.,* **29**(23), 4881–4891 http://www.lecb.ncifcrf.gov/~toms/paper/baseflip/.

45. Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA,* **73**, 804–808.

46. Escolar, L., de Lorenzo, V., and Perez-Martin, J. (1997) Metalloregulation in vitro of the aerobactin promoter of *Escherichia coli* by the Fur (ferric uptake regulation) protein. *Mol. Microbiol.,* **26**, 799–808.

47. Rudd, K. E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.,* **28**, 60–64.

48. Muller, K., Matzanke, B. F., Schunemann, V., Trautwein, A. X., and Hantke, K. (1998) FhuF, an iron-regulated protein of *Escherichia coli* with a new type of [2Fe-2S] center. *Eur. J. Biochem.,* **258**, 1001–1008.

49. Achenbach, L. A. and Genova, E. G. (1997) Transcriptional regulation of a second flavodoxin gene from *Klebsiella pneumoniae*. *Gene,* **194**, 235–240.

50. Robison, K., McGuire, A. M., and Church, G. M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.,* **284**, 241–254.

51. Lavrrar, J. L., Christoffersen, C. A., and McIntosh, M. A. (2002) Fur-DNA interactions at the bidirectional *fepDGC-entS* promoter region in *Escherichia coli* . *J. Mol. Biol.,* **322**, 983–995.

52. Baichoo, N. and Helmann, J. D. (2002) Recognition of DNA by Fur: a reinterpretation of the Fur box consensus sequence. *J. Bacteriol.,* **184**, 5826–5832.

53. Baichoo, N., Wang, T., Ye, R., and Helmann, J. D. (2002) Global analysis of the *Bacillus subtilis* Fur regulon and the iron starvation stimulon. *Mol. Microbiol.,* **45**, 1613–1629.

54. Fuangthong, M. and Helmann, J. D. (2003) Recognition of DNA by three ferric uptake regulator (Fur) homologs in *Bacillus subtilis*. *J Bacteriol,* **185**, 6348–6357.

55. Schneider, T. D. (2002) Consensus Sequence Zen. *Applied Bioinformatics,* **1**(3), 111–119 http://www.lecb.ncifcrf.gov/~toms/papers/zen/.

56. Schneider, T. D. and Stormo, G. D. (1989) Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res.,* **17**, 659–674.

57. Peck, L. J. and Wang, J. C. (1981) Sequence dependence of the helical repeat of DNA in solution. *Nature,* **292**, 375–378.

58. Rhodes, D. and Klug, A. (1981) Sequence-dependent helical periodicity of DNA. *Nature,* **292**, 378–380.

59. Head, C. G., Tardy, A., and Kenney, L. J. (1998) Relative binding affinities of OmpR and OmpR-phosphate at the *ompF* and *ompC* regulatory sites. *J. Mol. Biol.,* **281**, 857–870.

60. Lavrrar, J. L. and McIntosh, M. A. (2003) Architecture of a fur binding site: a comparative analysis. *J. Bacteriol.,* **185**, 2194–2202.

61. Hengen, P. N., Lyakhov, I. G., Stewart, L. E., and Schneider, T. D. (2003) Molecular flip-flops formed by overlapping Fis sites. *Nucleic Acids Res.,* **31**(22), 6663–6673.

62. Shultzaberger, R. K. and Schneider, T. D. (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.,* **27**(3), 882–887 http://www.lecb.ncifcrf.gov/~toms/paper/lrp/.

63. Kelley, R. L. and Yanofsky, C. (1982) Trp aporepressor production is controlled by autogenous regulation and inefficient translation. *Proc. Natl. Acad. Sci. USA,* **79**, 3120–3124.

Accession	Gene	Coordinate Range	#	R_i	%
U00096	*fepA-fes*	611808 – 611986	7	21.7	92
U00096	*fepB*	623839 – 624039	2	23.9	100
U00096	*entC*	624048 – 624096	6	27.6	139
U00096	*fur*	709863 – 710043	3	14.1	159
U00096	*tonB*	1308971 – 1309201	8	20.3	216
U00096	*cir*	2244950 – 2245050	5	20.8	113
U00096	*sodA*	4098230 – 4098420	4	21.2	95
U00096	*fecA*	4514280 – 4514340	3	19.3	97
U00096	*fecIR*	4515836 – 4515950	12	22.4	141
L01627	*hlyCABD*	100 – 700	31	9.3	96
M10930	*iucA1,2*	280 – 400	20	22.7	236

Table 1: Footprinted Fur binding sites in *E. coli*.

The footprinted regions were scanned with the Fur individual information model ($R_i > 0$) to test the model's validity. This table lists the GenBank accession number, the gene promoter that was footprinted, a range of the genome that includes the footprinted region and the entire cluster, the number of walkers in the region, the strongest R_i value in the same region, and the percent of the footprint covered by sequence walkers. Percentages greater than 100 indicate that the walker cluster covered the entire footprint as well as some of the flanking sequence on at least one end. In the case of *iucA*, two sites were used in the model (Fig. 1), although they are both part of a single contiguous footprint (7).

R_i	Genome Position	Coordinate	#
26.2	*fhuF*	4603345	11
24.5	*yoeA*	2066612	21
23.4	*ydiE*	1787605	14
22.6	*oppA*	1298970	5
21.8	*bfd*	3464671	11
21.4	*nohA*	1634627	5
21.4	*fepD – entS*[*ybdA*]	621440	7
20.9	*mntA*	2510785	9
20.6	*nohB*	579821	3
20.5	*yrhB*	3582383	16
20.3	*yojI*	2306703	7
20.1	*yddA*	1577361	20
20.0	*gpmA*	786856	5
19.2	*priC – ybaN*	490060	7
19.1	*yebN*	1903409	8
19.0	*ydhY*	1752759	23
18.9	*acnA*	1333487	10
18.8	*fiu*	840874	11
17.8	*ypjC – ygaQ*	2784037	23
17.8	*yohL – yohM*	2183927	10
17.8	*hyfA*	2599097	3
17.6	*yncD – yncE*	1521237	7
17.5	inside *gspC*	3453320	9
17.1	inside *yhaU*	3272486	16
16.7	inside *yahA*	331898	5
16.7	*yhhX – yhhY*	3578665	4
16.7	*fhuA*	167436	7
16.7	*exbB – metC*	3150072	3
16.6	*yahA*	331088	18
16.6	end of *sfmF* and *fimZ*	563069	14
16.5	*yceJ*	1118361	13
16.5	inside *fadD*	1886648	3
16.5	inside *elaD*	2381870	10
16.5	*appY*	582604	28
16.3	inside *yjbI*	4249428	16
16.3	inside *yjbI*	4249205	26
16.3	*yqjH*	3214194	12
16.2	inside *yjgL*	4474598	27
16.2	inside *fliC*	2000942	7
16.1	inside *ybiD*	3787592	5

Table 2: 40 predicted Fur binding regions in the *E. coli* genome.

The strongest individual information value (R_i, bits) in each region is shown, followed by the region's position in the genome (47), the center coordinate of the ± 200 bases that were scanned, and the number of sequence walkers ($R_i > 0$) in the region.

```
                                    ---                   +++
                                    111---------  +++++++++111
                                    210987654321012345678901 2
                                    ......................... bits
fepA-fes U00096  611892 +  1 aatgcaaatagttatcaataatatt 21.7
         U00096  611892 -  2 aatattattgataactatttgcatt 21.7
fepB     U00096  623954 +  3 atcaataatgcttctcattttcatt 23.9
         U00096  623954 -  4 aatgaaaatgagaagcattattgat 23.9
entC     U00096  624069 +  5 aatataaatgataatcattattaaa 27.6
         U00096  624069 -  6 tttaataatgattatcatttatatt 27.6
fur      U00096  709943 +  7 tgagataatgcgtatcattatagaa 14.1
         U00096  709943 -  8 ttctataatgatacgcattatctca 14.1
tonB     U00096 1309056 +  9 attgaatatgattgctatttgcatt 20.3
         U00096 1309056 - 10 aatgcaaatagcaatcatattcaat 20.3
cir      U00096 2244975 + 11 aatgcaaacgataacaattatcaat 20.8
         U00096 2244975 - 12 attgataattgttatcgtttgcatt 20.8
sodA     U00096 4098301 + 13 tacggcattgataatcattttcaat 21.2
         U00096 4098301 - 14 attgaaaatgattatcaatgccgta 21.2
fecA     U00096 4514321 + 15 aatcgaaataagaattattttcctt 19.3
         U00096 4514321 - 16 aaggaaaataattcttatttcgatt 19.3
fecIR    U00096 4515857 + 17 tatgagaatggttatcattacaatt 22.4
         U00096 4515857 - 18 aattgtaatgataaccattctcata 22.4
hly      L01627     547 + 19 actacaatatataatcataaagaaa  6.7
         L01627     547 - 20 tttctttatgattatatattgtagt  6.7
iucA1    M10930     358 + 21 attgataatgagaatcattattgac 22.7
         M10930     358 - 22 gtcaataatgattctcattatcaat 22.7
iucA2    M10930     376 + 23 tattgacataattgttattatttta 12.9
         M10930     376 - 24 taaaataataacaattatgtcaata 12.9
```

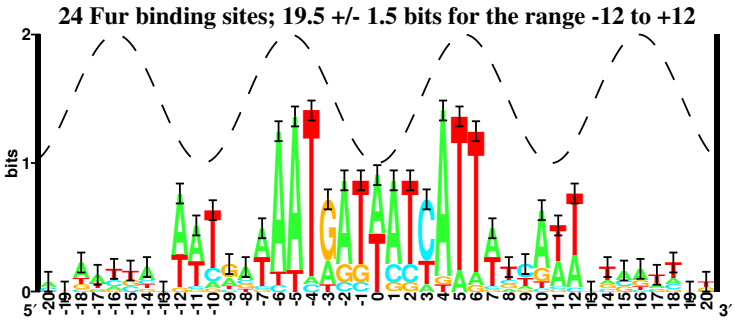

Figure 1: Aligned listing of *E. coli* Fur binding sequences and sequence logo. This is the optimum alignment and resulting sequence logo of DNase I footprinted sequences and their complements, corresponding to alignment 1 in Fig. 2. The numbers along the top of the sequences are read vertically and denote the distance from the center of the aligned sequences. The listing is of the strong conservation seen in the logo below, from base −12 (on the left) to +12 (on the right); this conserved region, with $R_{sequence} = 19.5 \pm 1.5$ bits (15), was used as the scanning model. Each line contains the genetic region in *E. coli*, GenBank accession number, coordinate of the zero base, orientation of the sequence fragment relative to the GenBank sequence, the sequence number, the sequence used in the creation of the logo, and individual information in bits. The sequence logo was derived from the 12 *E. coli* Fur binding sites and their complements as listed. The sine wave represents the twist of B-form DNA, 10.6 bases per turn (27).

(A)

```
Alignment 1, 296 occurrences, relative aligned bases:
0  0   0  0   0 0   0  0   0 0   0 0   0  0   0  0   0  0   0  0   0 0   0 0
Alignment 2, 38 occurrences, relative aligned bases:
6 -6   0  0   0 0   0  0   0 0  -6 6   6 -6   0  0   0  0  -6  6   0 0   0 0
Alignment 3, 2 occurrences, relative aligned bases:
6 -6   0  0   0 0   0  0  -6 6  -6 6   6 -6   6 -6   0  0   6 -6   0 0  -6 6
Alignment 4, 126 occurrences, relative aligned bases:
3 -3   3 -3  -3 3   3 -3  -3 3  -3 3   3 -3  -3  3   3 -3   3 -3  -3 3  -3 3
```

(B)

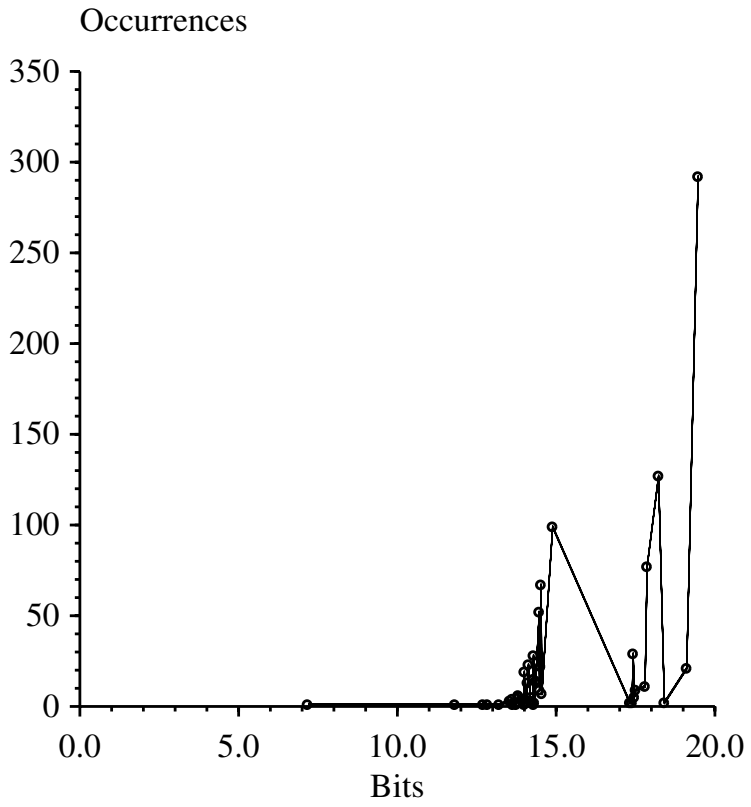

Figure 2: Alignments created by the **malign** program.
(A) Realignment vectors for the best four alignments show sequences shifting by 6 base pairs. Each number represents the distance that a sequence and its complement were shifted from the original alignment. (B) Each information content in this graph represents one unique alignment. The information content of each alignment is plotted against the number of times that that particular alignment occurred during 1000 realignments. The best alignment (number 1) had 19.5 bits and occurred 296 times.

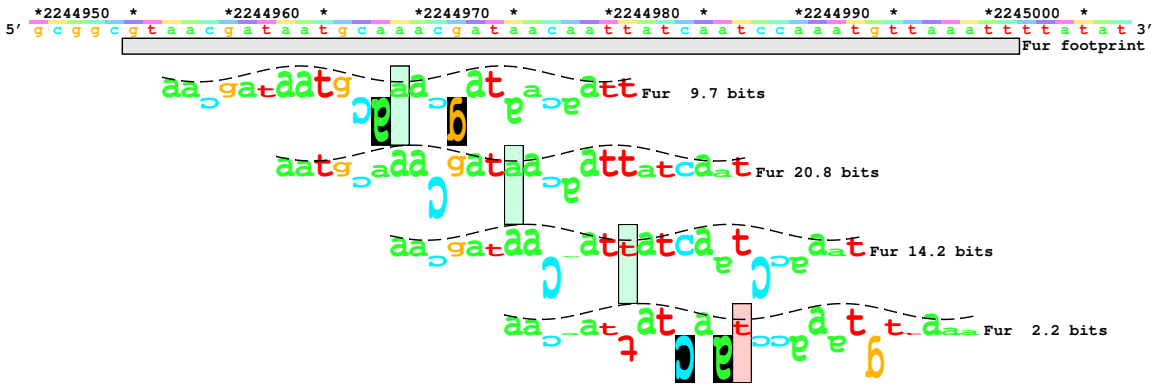

Figure 3: *cir* promoter scan displaying 6-base spacing of multiple Fur walkers. Four sequence walkers were found in the *cir* promoter region with $R_i > 0$ bits. The gray bar below the DNA sequence marks the DNase I footprint in the region (28). Bases that are not in the Fur model have a black background. The six-base spacing between walkers can be seen using the color strip above the sequence. This strip was produced using the **live** program, set to run through the rainbow every 6 bases. The color yellow is located above the zero base of each site (vertical bars). The walker with a center bar of pink has an $R_i < 3$ bits; those with $R_i \geq 3$ bits are green.

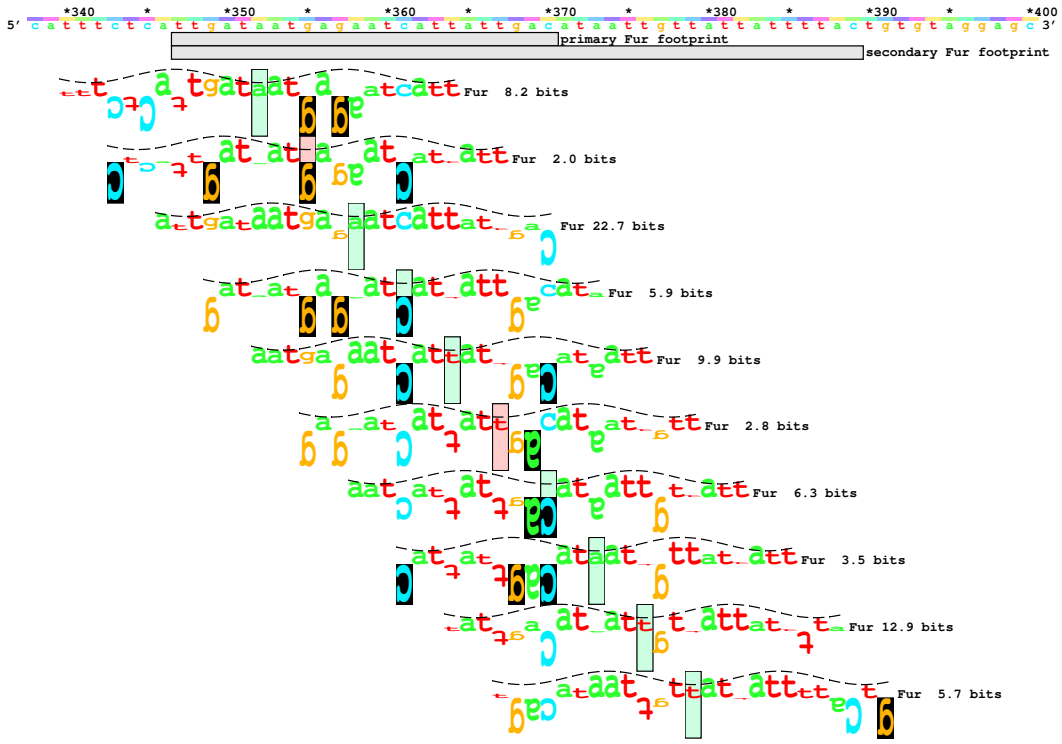

Figure 4: *iucA* promoter scan displaying primary and secondary Fur footprints. With low concentrations of Fur the primary footprint is observed, while at higher concentrations the secondary footprint appears (30). The walkers correspond to the longer secondary footprint well, supporting the hypothesis that Fur binds at successive sites on the DNA which are separated by six and three bases. Note that the walker with the highest information content (22.7 bits, 358) is in the primary footprint, while a weaker walker (12.9 bits, 376) overlaps the secondary footprint.

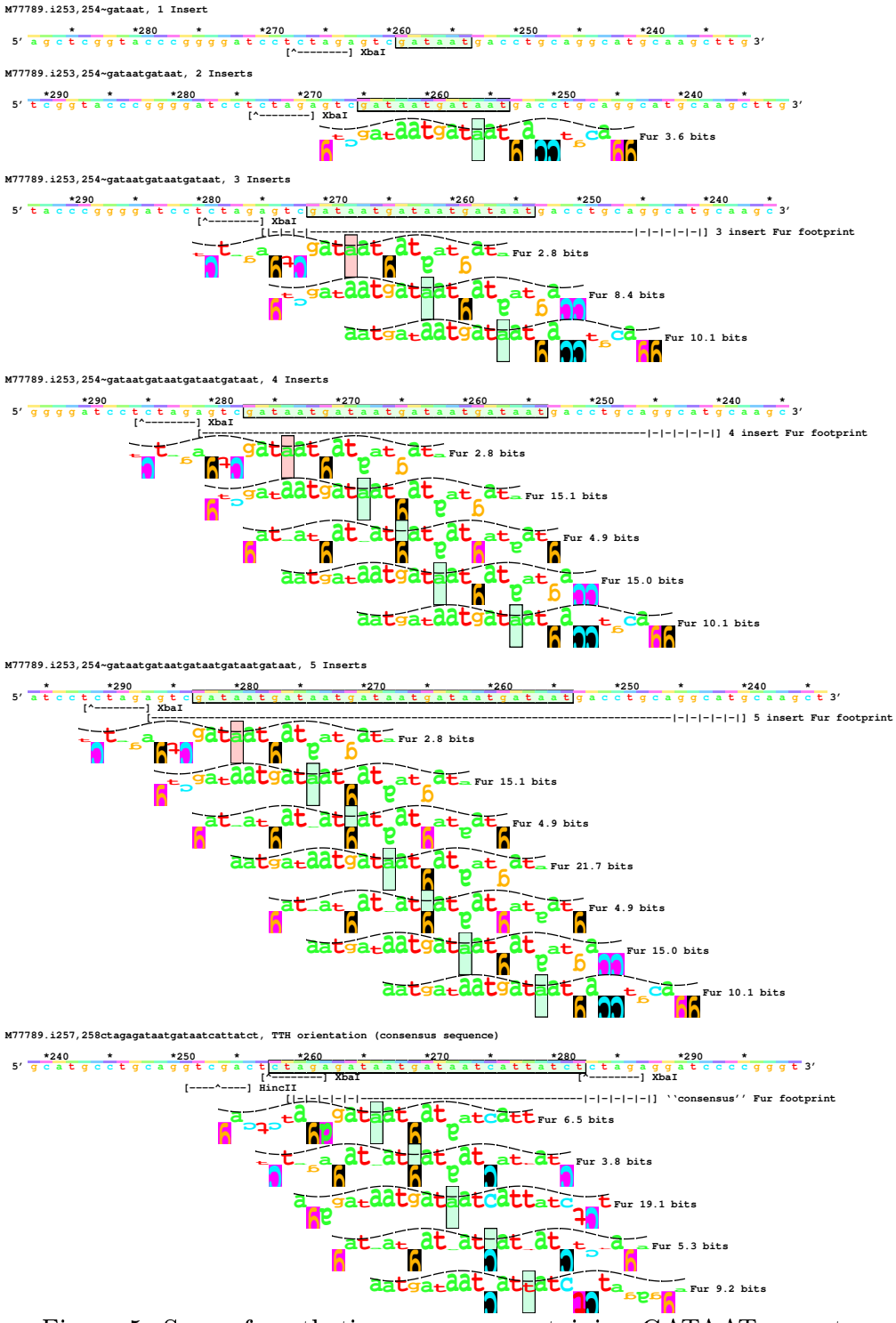

Figure 5: Scan of synthetic sequences containing GATAAT repeats. pUC19 (GenBank accession M77789) derivatives containing inserts with GATAAT repeats (boxed) were footprinted (5). When scanned with the Fur weight matrix ($R_i > 0$), sequence walkers appear coinciding with the documented footprints and a weak interaction for the 2-insert case. The dashed lines below the sequence represent restriction sites and Fur footprints; the hatched lines inside the footprints represent faint protection by Fur. To keep the figure a manageable size, the walker lower bounds are at -2 bits. Bases that go below -2 bits are surrounded by a pink box.

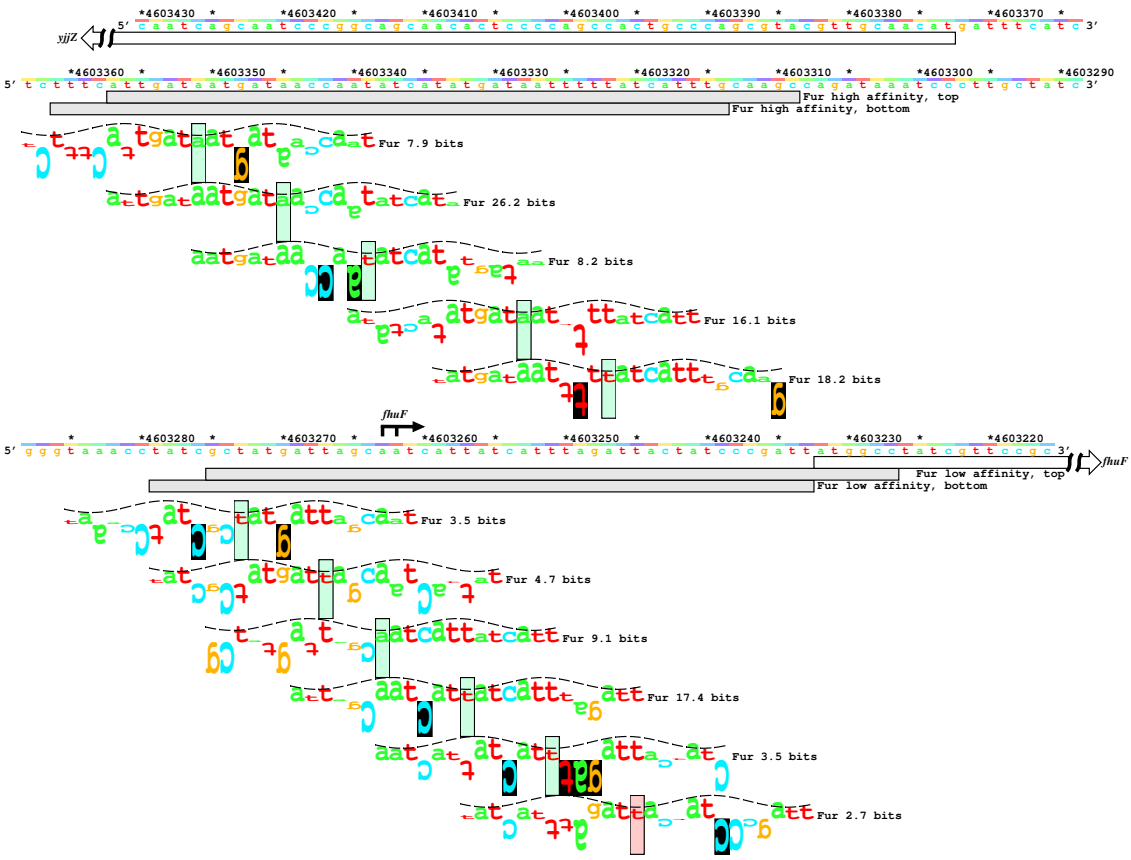

Figure 6: Scan of the *fhuF* promoter region for Fur sites.
The promoter region of *fhuF* contains the strongest individual information site in the entire *E. coli* genome (Table 2). In this region there are two clusters of Fur sequence walkers ($R_i > 0$). When footprinted (Fig. 7), sequences protected from DNase I digestion (gray bars) corresponded closely to the walkers. Two distinct segments were protected by Fur; at minimal concentrations of Fur, only the segment marked 'high-affinity' was protected. Higher concentrations of protein also protected the 'low-affinity' segment downstream, closer to the *fhuF* translational start. The solid black arrows above the DNA at coordinates 4603263 and 4603262 indicates the *fhuF* transcription starts. The open arrows at coordinates 4603373 and 4603232 indicate the *yjjZ* and *fhuF* translation starts respectively.

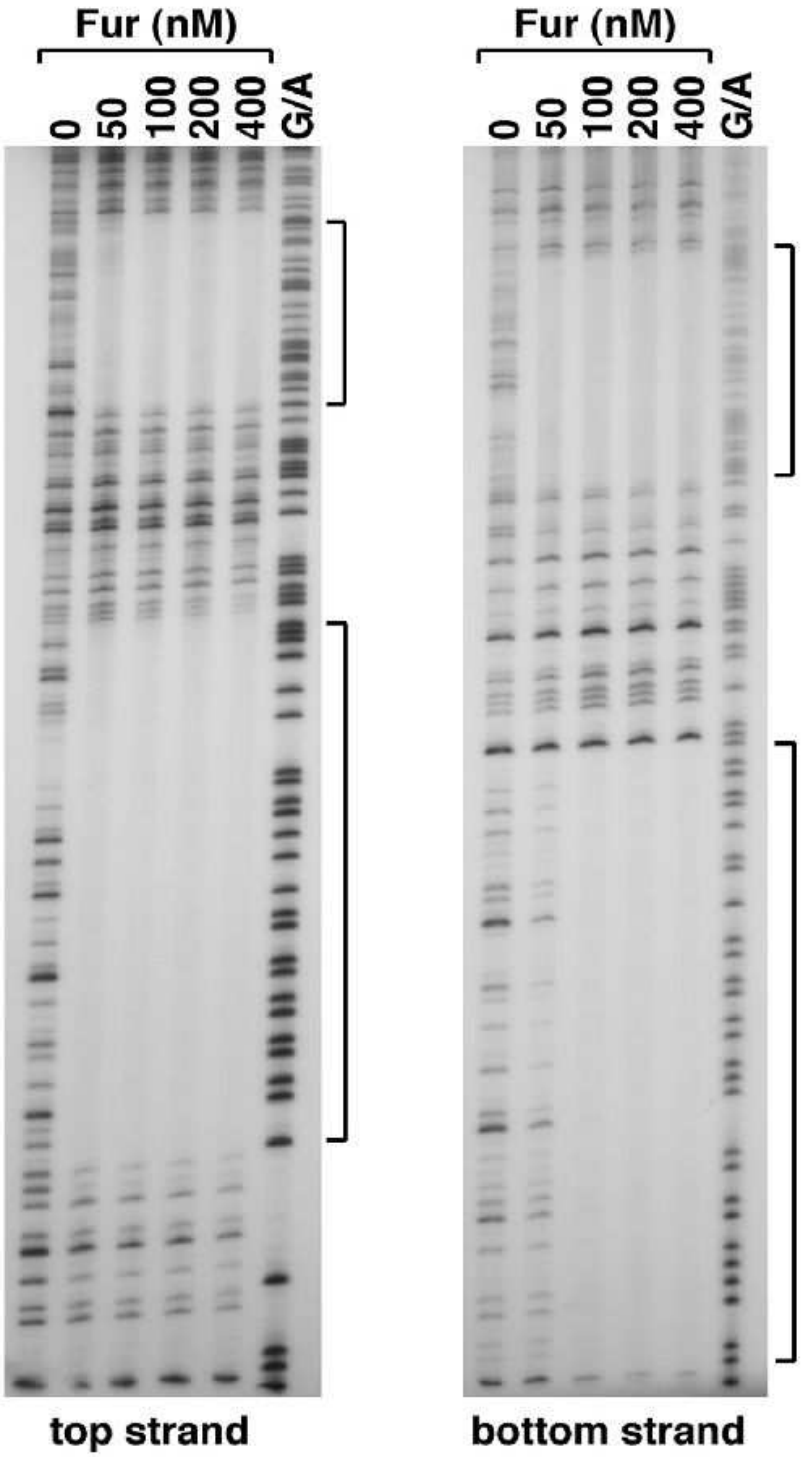

Figure 7: DNase I footprints of the *fhuF* promoter region.
DNase I footprinting on the *fhuF* promoter by Fur showed two regions protected by the protein, marked in the figure by brackets. The footprinting samples were run in parallel with Maxam-Gilbert sequencing ladders (marked by G/A). The corresponding sequences are shown in Figure 6.

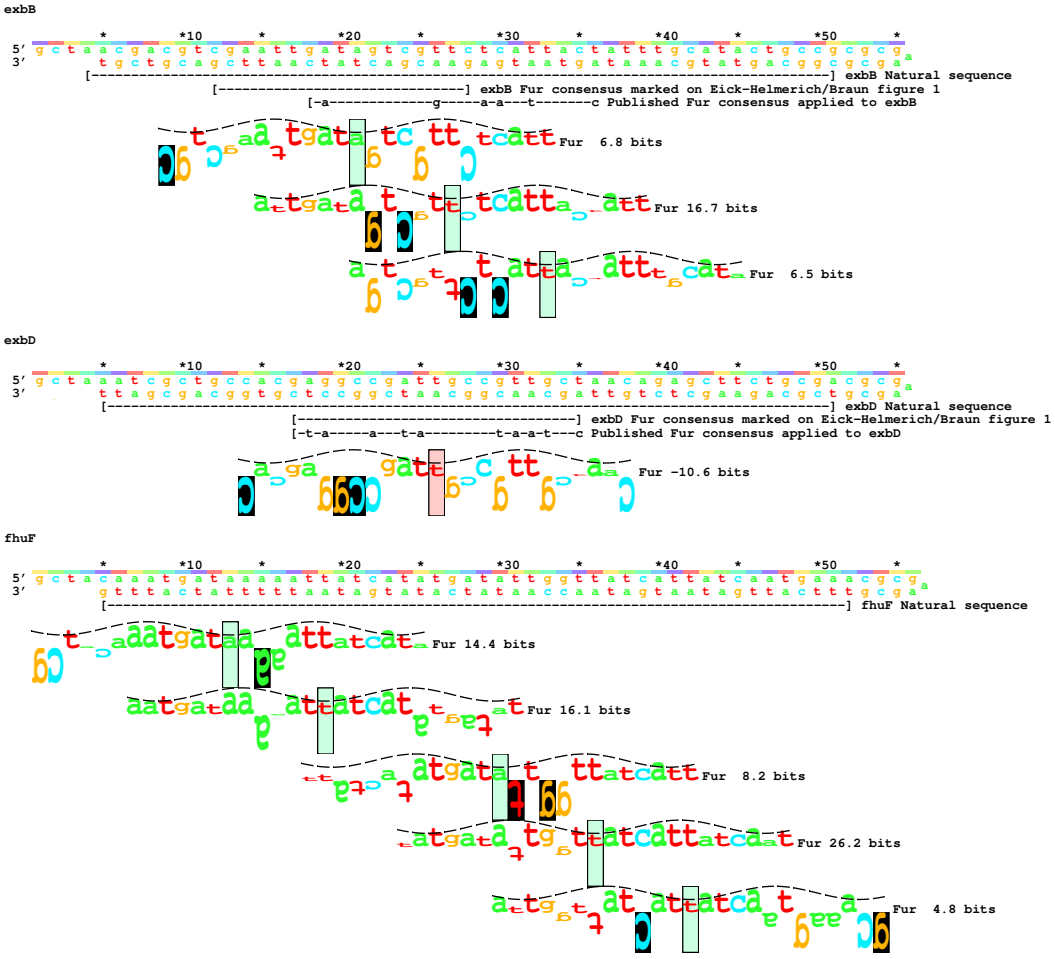

Figure 8: Oligonucleotides used in gel binding assays for *exbB*, *exbD*, and *fhuF* (Fig. 9). Portions of the *exbB* and *exbD* promoter regions that had been predicted to bind Fur based on a consensus sequence (37) were incorporated into oligonucleotides containing hairpins (23). The region of the *fhuF* promoter that bound Fur with high affinity (Fig. 6) was also incorporated into an oligonucleotide as a positive control. The natural sequences of each promoter are marked by dashes (*E. coli* sequences: *exbB*, 3150049 to 3150096; *exbD*, 3149433 to 3149479; *fhuF*, 4603314 to 4603361). The Fur consensus sequences, as proposed by Eick-Helmerich and Braun and as found by the Delila **search** program, are marked with mismatches indicated. Scans were with $R_i > 0$ for *exbB* and *fhuF*; $R_i > -11$ for *exbD*. Walkers with a center bar of pink have an $R_i < 3$ bits; those with $R_i \geq 3$ are green.

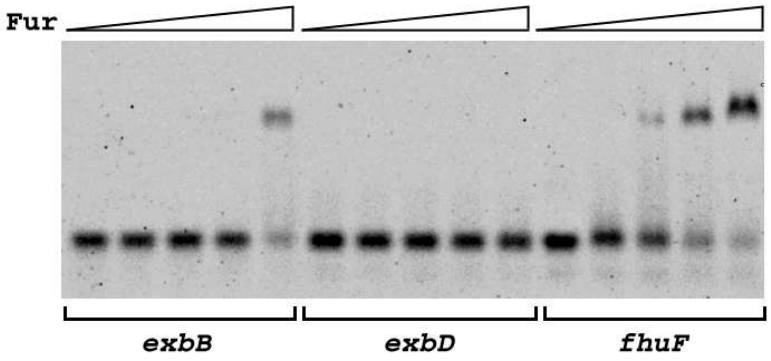

Figure 9: Gel mobility shift assay for *exbB*, *exbD*, and *fhuF* oligonucleotides.
This gel mobility shift assay displays shifts in the *exbB* and *fhuF* but not *exbD* oligonucleotides (Fig. 8) when incubated with Fur protein. The hairpin oligonucleotides in each lane are indicated by brackets on the bottom of the gel, and increasing protein concentration (0, 80, 160, 320, and 640 nM) is indicated by triangles above the gel.

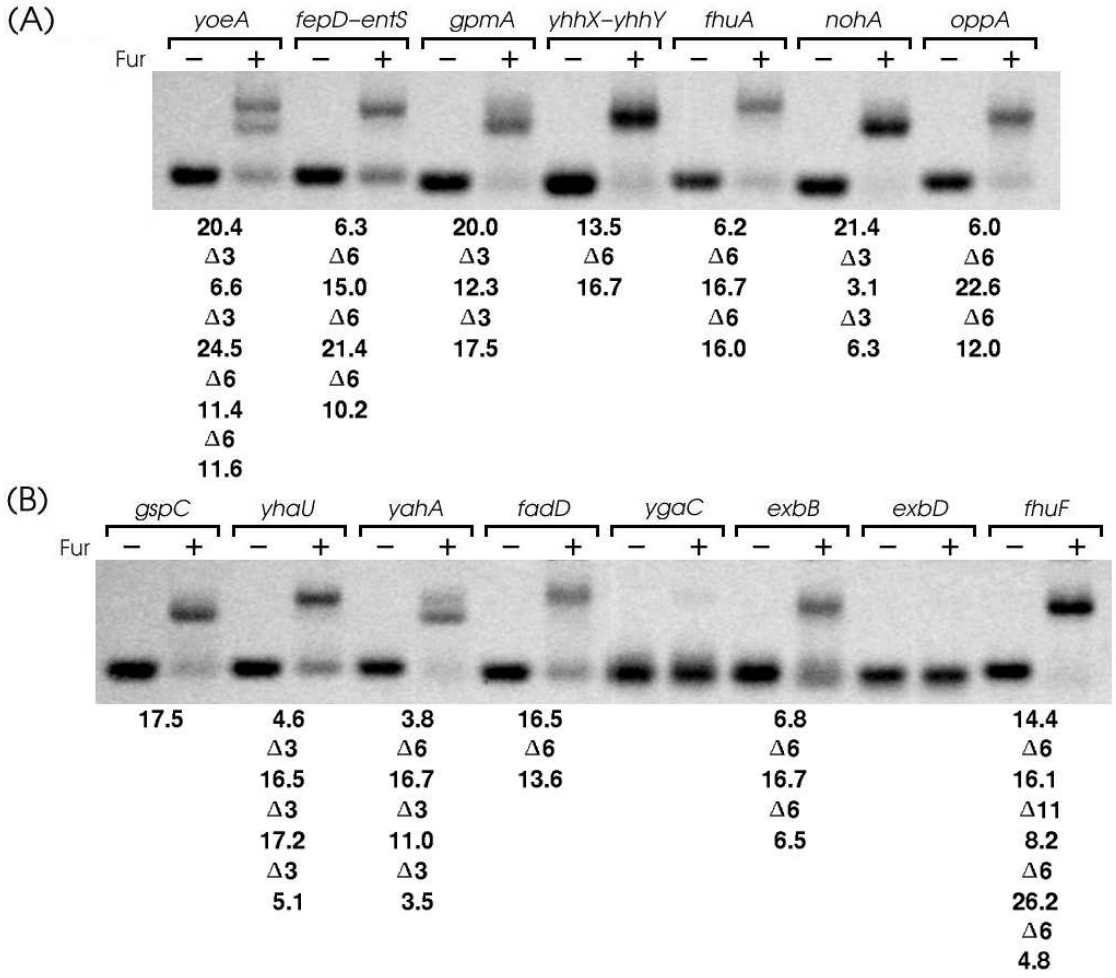

Figure 10: Gel mobility shift assay for additional oligonucleotides.
Oligonucleotides containing predicted Fur binding sites were incubated without (-) or with (+) 150 nM Fur protein and gel electrophoresed to further test the model. Below each set of lanes is the strength in bits of the sequence walkers found on each oligo (numbers only) and the number of bases separating the zero coordinates of the walkers (numbers preceded by a delta). (A) The first set of oligos contain predicted Fur sites that were found using both the Fur and promoter information theory models (*yoeA*, *fepD-entS* [formerly *ybdA*]; *gpmA*, *yhhX-yhhY*, *fhuA*, *nohA*, and *oppA*). (B) The second set of oligos contain predicted Fur sites located within genes (*gspC*, *yhaU*, *yahA*, and *fadD*) and a consensus-predicted site that does not contain an sequence walker (*ygaC*). The *exbB*, *exbD*, and *fhuF* oligos from the previous gel shift (Fig. 9) were also included as controls. As expected, all oligos that contain sequence walkers exhibit one or more mobility shifts following incubation with 150 nM Fur; the *ygaC* and *exbD* oligos, which do not have sequence walkers, do not shift.